DOCUMENT RESUME

ED 319 791 TM 015 052

AUTHOR Legg, Sue M.; Buhr, Dianne C.
TITLE Investigating Differences in Mean Score on Adaptive and Paper and Pencil Versions of the College Level Academic Skills Reading Test.
PUB DATE Apr 90
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *College Entrance Examinations; Community Colleges; Comparative Testing; *Computer Assisted Testing; Error of Measurement; Field Studies; Followup Studies; Higher Education; *Reading Tests; Scores; Student Motivation; *Test Format; Timed Tests; Undergraduate Students
IDENTIFIERS *College Level Academic Skills Test; *Paper and Pencil Tests

ABSTRACT

Possible causes of a 16-point mean score increase for the computer adaptive form of the College Level Academic Skills Test (CLAST) in reading over the paper-and-pencil test (PPT) in reading are examined. The adaptive form of the CLAST was used in a state-wide field test in which reading, writing, and computation scores for approximately 1,000 students were compared for the March or June 1988 administrations of the PPT and the spring of 1988 administration of the computer adaptive test (CAT) version. The field study data were analyzed to address measurement error, content coverage and reading load, and student motivation. A follow-up study used data from 361 community college freshmen and sophomore students who had taken the October 1989 or March 1990 paper-and-pencil version of the CLAST but had not yet received their scores; these subjects were administered the Nelson-Denny Reading Test, as well as the CAT form of the CLAST. The follow-up study failed to replicate the difference in mean scores for the computer adaptive and paper-and-pencil versions of the reading test that were found in the preliminary analysis. Overall results indicate that students should be allowed a longer testing period for the CAT version of the CLAST than for the PPT version. Seven data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319791

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SUE M. LEGG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Investigating Differences in Mean Score on Adaptive and Paper and Pencil Versions of the College Level Academic Skills Reading Test

by

Sue M. Legg, Ph.D. and Dianne C. Buhr, Ph.D

University of Florida

Paper Presented at the Annual Meeting of the National Council of Measurement in Education

Boston, Massachusetts April 1990

TM 015052

BEST COPY AVAILABLE

Investigating Differences in Mean Score on Adaptive and Paper and Pencil Versions of the College Level Academic Skills Reading Test

Objectives of Inquiry

The purpose of this study is to examine the possible causes of a sixteen point mean score increase for the computer adaptive form of the College Level Academic Skills Test (CLAST) in reading over the paper and pencil test (PPT) in reading. The adaptive form of the CLAST was used in a state-wide field test in which reading, writing and computation scores for approximately 1,000 students were compared for the March or June 1988 administrations of the PPT and computer adaptive test (CAT) version administered in the spring of 1988.

Mean scores for the writing and computation tests were nearly identical on the two versions, and they approximated the mean scores for the state population. CAT reading test scores were significantly higher than the PPT scores for the same students; however, the mean PPT test scores in reading for the sample were some what lower than the mean reading scores for the state population (State X = 313 and sample X PPT = 306). The results of the field test analysis indicated that score differences were not due to factors such as racial/ethnic status, age, computer experience, or sex.

Possible explanations for the reading score differences were developed with the assistance of the consultants to the project, Dr. David Miller and Dr. Mark Reckase. Each of these explanations was examined, leading to the hypothesis that timing and format were the most likely factors to increase the CAT reading scores. A follow up study was then conducted to investigate possible effects of timing and format on the reading scores.

This paper reviews the preliminary analyses from the field test that led to the hypothesis about time and format and presents the findings of that follow up study. The following issues are addressed using data from the field test:

1) Measurement error
2) Content coverage and reading load
3) Student motivation

Instruments

The CLAST consists of the three objective tests in reading, writing and mathematics and an essay. It is administered to college sophomores as part of the requirement for upper division status. The PPT has 40 multiple choice items in writing, 41 in reading, and 55 in mathematics. The CAT item banks contain the same proportion of items in each content area as represented on the PPT. Approximately one half of the number of PPT items are administered on the CAT. The reading test consists of 200 to 400 word passages followed by one or more items. The passages were too long to fit on the computer screen; thus, the CAT required students to scroll through the passages by pressing

the up or down arrow keys. Items were presented one at a time with the passage in a window at the top of the screen.

The Nelson Denny Reading Comprehension test was administered to students as part of the follow up study on the effects of timing and format on the CAT reading subtest.

Methodology and Results of Preliminary Analyses

Measurement Error

The procedures for building the item banks were examined to identify possible sources of measurement error. Items were selected using the latest difficulty estimates, but the item base values were used for scaling. If shifts in difficulty had occurred, then the score scale would be affected. The PPT was rescored using the current difficulty estimates, and a comparison of the old and new difficulty values showed a difference of .01 logits and essentially no difference in the two score scales. The correlation between the scales was .95 and sample mean scores were 305.92 and 305.51.

If the precision of measurement of the CAT and the PPT versions is focused at different ability levels, then different score distributions can result. While reliability at the cut score was high, there was no comparison of the test characteristic curve for the PPT and the average test characteristic curve for the CAT. However, for the CAT Reading test, the average number of items per examinee was 20 of the 109 available items. A few items were heavily used for the CAT test, while others were selected for fewer than 10 of the 545 students. The large number of empty cells in the data matrix did not allow for reestimation of the item parameters for the CAT data.

Inaccuracies in modeling might cause some of the differences in the score distributions. The predicted proportion of correct response was compared to the actual proportion of correct response for items at six ability levels. The item fit statistics were also examined. Only one of the 36 scored items showed a between fit statistic of greater than 3.00, the criterion used for misfit for the paper and pencil test.

Content and Reading Load

If the tests measure more than one dimension, and the test versions emphasize different dimensions, different score distributions could result. The CAT reading bank contained the same proportion of items in each content area as represented on the PPT. A factor analysis of the PPT version was conducted early in the testing program, and a dominant factor was found. Given the large number of empty cells in the CAT data matrix, it was not possible to calculate the tetrachoric correlations for the factor analysis.

If the CAT had fewer words per item, then the differences in score might be explained by differences in reading load. Word counts were completed for all passages in the CAT item bank and on the PPT test. Standard words per passage were calculated and compared. The CAT version contained a greater number of standard words per passage than

the PPT; the number of standard words per passage varied from 101.33 to 464.17 for the CAT and from 93.50 to 478 for the regular examinations. Average number of standard words for the CAT was 288.52 and was 225.53 and 280.38 for the two regular administrations. From four to five passages and 19-24 items were administered to each examinee on the CAT and 11-13 passages with 44 items appear on the PPT. Differential reading level of the tests does not appear to be a factor which would result in higher scores for the CAT.

Differential Student Motivation

Students were awarded the higher of the two scores from the CAT and the PPT. Administration of the CAT was balanced prior to and after the PPT administration. While no differences were found for administration effects, it was possible that students were less likely to be motivated for the PPT. The records of responding for all examinees on the PPT were printed and examined. Records in which four or more items were omitted in a row or that contained obvious random response patterns such as a series of the same response were dropped from the data, and the mean scores were recalculated. Only a small increase in mean scores was found. Score differences on the CAT and PPT in reading were not replicated on the mathematics and writing tests, which buttresses the argument that motivation was not a factor in the CAT score increase.

Timing and Format Follow-Up Study

The PPT reading test is administered following the writing test with a single time limit. While the PPT allows ample time, students taking the CAT may have felt less restricted by time. Average time per item on the CAT was 1.45 minutes, while the average time per item on the PPT was .99 minutes. The correlation between the time used per item on the CAT and score on the PPT was - .32, thus indicating that low score on the PPT test was associated with longer time on the CAT. Average time per CAT item for each of ten PPT score groups, and the relationship between PPT score groups and average time used in five categories of time on the CAT support the contention that students with low PPT test scores in reading use more time per item on the CAT. This analysis indicated that time used on the CAT may have contributed to the increase in reading scores.

The possibility that time and format effects combined to increase the reading scores was investigated by administering the CAT reading test under two time and two format conditions. For one group of students, the time allowed was the current time limit (55 minutes) for the PPT CLAST reading test. For the second group, the time allotted was 45% of the regular time (25 minutes), since the average number of items represents 45% of the items on the PPT.

In order to separate out format effects from time effects, two versions of the CAT Reading test were administered, one to replicate the previous field test and one that presented the CAT items on the computer screen but directed students to read the passages from a hard

copy booklet. Time was fully crossed with format, so that there were four methods of CAT administration.

Data also included scores from the Nelson Denny Reading Comprehension Test used as a covariate in a repeated measure analysis of covariance and responses from a questionnaire administered to students to assess their testing strategies relating to time and format.

Support for a format effect due to presentation of text on computers was reported by Mason (1987) in The Relationship between Computer Technology and the Reading Process: Match or Misfit? and by Keene and Davey (1987) in their article Effects of Computer-Presented Text on LD Adolescents' Reading Behaviors. It may be that the increase in time was due to line by line focusing on text by lower scoring students. This focus on individual lines may account for the increased time and may also have improved reading comprehension resulting in higher scores. If this is the case, there are obvious instructional and assessment implications.

Sample. The sample for the follow-up study consisted of community college freshmen and sophomore students who had taken the October 1989 or March 1990 PPT version of the CLAST but had not yet received their scores. The sites were selected to include a representative number of minority students in the sample. Table 1 shows the breakdown of the sample by ethnic group, as compared with the total population for the PPT administrations.

Table 1. Classification by Ethnic Group

	Percentage of	
Ethnic Group	Sample	Total
White	58%	64%
Black	14%	14%
Hispanic	18%	14%
Asian	5%	4%
Non-resident alien	3%	3%
Other	2%	1%

Procedures. Students were first administered the reading comprehension portion of the Nelson-Denny Reading Test (Brown, Bennett, and Hanna, 1981). This test consisted of eight reading passages, with a total of 36 multiple-choice questions, with a time limit of 20 minutes. Alternate-forms reliability was reported as .81 for grade 14.

Students were assigned randomly to a computer that had the directions for taking the CAT on the screen. The number of students assigned to each of the groups was balanced, with 89 examinees in Group I, 91 examinees in Group II, 91 examinees in Group III, and 90 examinees in Group IV.

Table 2 illustrates the four methods of CAT administration.

Table 2. Design for the Experiment

	Format	
Time	Passages on screen	Passages on paper
55 min	I	III
25 min	II	IV

After finishing the CAT, students were asked to respond to a short questionnaire assessing their reactions to the computerized testing format. In addition, specific questions about their testing strategies relating to time usage and format constraints were included. See the Appendix for a copy of the questionnaire.

Data Analysis. An analysis of covariance (Ancova) with score on the CLAST as the repeated measure, score on the Nelson-Denny as the covariate, and test time/format group as the independent variable was conducted to answer the primary question: Is the difference between the CAT and the PPT, controlling for initial ability, affected by testing time and/or format?

Results and Discussion: Test-Taking Strategies. Seventy-two percent of the students reading the passage from the screen responded that scrolling helped them to focus their attention on the relevant part of the reading passage, and 82% indicated no difficulty in locating the part of the passage needed to answer the question. Sixty-five percent of these students indicated that they scrolled several lines at a time, while 34% indicated that they scrolled one line at a time.

Interestingly, a much higher proportion of those students reading from the screen indicated some eye strain by the end of testing (42% versus 25%). Frequencies of response to the questionnaire items for the two format groups are given in Appendix B.

Results and Discussion: Time Differences. There was a significant difference in testing time, $F(3,358) = 13.93$, $p=.0001$, with the average testing time across groups at 23 minutes. The average number of minutes for the total test for each of the groups, along with the average number of minutes per item are given in Table 3. While it may be expected that Groups I and III would use more time than the other groups, it was somewhat surprising that the average amount of time used for all groups differed by only four minutes. This difference, however, appeared to be related to the students' scores.

Table 3. Test Time for Each of the Groups

Group	Average Test Time	Average Min/Item	N
I	24.77	1.24	88
II	20.73	1.06	88
III	24.75	1.22	89
IV	21.31	1.08	88

Note: Time data were missing for eight examinees.

In contrast to the field test study of the CAT CLAST examination, scores on the CAT and PPT versions did not differ significantly for this group of students ($t = -3.10$ (p .01)). The CAT CLAST mean score was 317.44, while the PPT mean score was 321.41. Average score on the Nelson-Denny Reading Test was 23.00. Table 4 shows the correlations of Nelson-Denny, CAT CLAST, PPT CLAST, and testing time on the CAT, for the total group of 361 examinees.

Table 4. Correlation of Testing Time, Nelson-Denny, CAT and PPT CLAST

	CAT CLAST	PPT CLAST	TEST TIME
Nelson-Denny	.56	.60	- .49
CAT CLAST		.68	- .08
PPT CLAST			- .26

Students in Group I scored highest on the CAT (see Table 5). This was the only group for which mean score on the CAT CLAST was higher than mean score on the PPT CLAST. For Group I students, mean score on the CAT CLAST was about two points higher than that for the PPT CLAST. However, for the other groups, mean score on the CAT CLAST was from three to eight points lower than that for the PPT CLAST.

Table 5. Scores on the Nelson-Denny, CAT, and PPT for the Four Groups

Group	Nelson-Denny	CAT CLAST	PPT CLAST
I	22.54	321.04	319.28
II	23.65	312.97	320.38
III	21.35	314.95	318.27
IV	24.04	319.65	327.61

Results and Discussion: Analysis of Covariance. There was a significant difference due to test time/format group, $F(3,357) = 3.41$,

$p = .02$. The adjusted mean CAT score for students in Group I, who read the passages from the screen and had 55 minutes for testing, was significantly higher than those for Groups II and IV, who had only 25 minutes for testing. Although the difference was not significant, the adjusted mean CAT score for students in Group I was also about four points higher than that for Group III, who also had 55 minutes but read the passages from a test booklet. Mean testing time and adjusted mean CAT scores are reported in Table 6. It appears that the higher scores are directly related to the testing time, even though the average time taken for all four groups was less than the allotted time for Groups II and IV. Mean time per item was 1.23 minutes for Groups I and III and 1.07 minutes for Groups II and IV, which indicated that reading from the screen took more time than locating and reading a passage in a test booklet. The additional time was beneficial to the examinees.

Table 6. Adjusted Mean CAT Scores for Each Test Time/Format Group

	Group			
	I	II	III	IV
	55 min/Screen	25 min/Screen	55 min/Booklet	25 min/Booklet
X Score	322.67	312.44	318.80	314.69
X Time	24.77	20.73	24.75	21.31

Distributions of CAT scores by ability group were examined to see if the time/format differences were greater for lower ability than for higher ability students. Field test results indicated that as expected lower ability students took more time per item, but our hypothesis for this study was that lower ability students might also focus on the reading task better when they used the scrolling function on the computer screen. This ability to focus line by line could result in higher scores for students.

Conclusions

The impetus for this study was the need to explore the differences for the computer adaptive and regular versions of the CLAST reading test. The preliminary analysis of the results of the field test data indicated that time and format were the most likely causes of the differences in scores. The follow-up study investigated the differences in CAT scores due to time and format but failed to replicate the difference in mean scores for the CAT and regular versions of the reading test. Significant differences were found due to the effect of time on the CAT; more time was related to an increase in scores for students who read the passages from the screen (Group I). Scores were higher for Group III students who took the adaptive version but read the passages from a booklet than the scores for students in the two groups with the shorter time difference. This difference, however, was not significant.

The failure to replicate the difference in scores for the regular and adaptive version is difficult to explain. Students in the original field test sat both the reading and writing tests but sat the CLAST reading test and the comprehension portion of the Nelson Denny in the follow-up study. It is conceivable that students paced themselves differently in the follow-up study even though there was no apparent time pressure to complete the examination, and students did not truncate the adaptive test in the follow-up study.

The explanation that more time increases scores may be too simple. The difference in mean time actually used for the two time groups was about four minutes. The groups that had 55 minutes used an average of 24.75 minutes while the groups with 25 minutes used about 21 minutes. It is likely that reading from the computer screen simply takes longer than reading from the booklet. While the group that had both the longer time and read the passages from the screen received the highest mean scores, these scores would not increase if testing time were increased; these students did not use the 55 minutes allotted. The implication for test administration is that students should be allowed the longer testing period for the CAT version in order to more closely replicate the paper and pencil test score.

Frequency of Response to Each Questionnaire Item
For Each Format Group

Item	Format Group	Response SA	A	D	SD	N
Procedures easy to follow	Screen	.72	.25	.03	.00	184
	Booklet	.74	.25	.00	.00	184
Enough practice responding	Screen	.50	.43	.04	.03	184
	Booklet	.56	.37	.06	.01	183
No difficulty pacing self	Screen	.35	.43	.16	.06	184
	Booklet	.31	.42	.22	.05	182
Felt eye strain	Screen	.10	.32	.40	.17	183
	Booklet	.01	.19	.48	.32	183
Had problems scrolling	Screen	.05	.15	.51	.29	182
Scrolling helped focus	Screen	.27	.47	.21	.05	184
No problem locating part of passage	Screen	.37	.46	.15	.02	184
	Booklet	.39	.48	.09	.04	183
Difficulty finding passage	Booklet	.04	.04	.37	.55	184
Reading comprehension affected	Screen	.15	.28	.36	.21	182
Concentration affected	Screen	.15	.32	.36	.17	184
Had trouble reading due to						
a. glare	Screen	.04	.11	.48	.37	170
b. brightness	Screen	.00	.04	.53	.43	170
c. letters too close	Screen	.03	.12	.46	.39	174
d. lines too close	Screen	.05	.20	.39	.35	173
Was more anxious with CAT	Screen	.16	.31	.33	.20	183
	Booklet	.17	.27	.26	.31	183
Prefer CAT over PPT	Screen	.34	.29	.29	.08	178
	Booklet	.42	.30	.21	.07	184

Note. SA=Strongly Agree, A=Agree, D=Disagree, SD=Strongly Disagree

Sue Legg, Ph.D., University of Florida
Dianne C. Buhr, Ph.D., University of Florida
10. Computer Applications

Abstract

This study examines the possible causes of a mean score increase for the computer adaptive form of a college level reading test over the traditional paper and pencil version. Issues addressed in the study include timing and format, measurement model procedures, content coverage and reading load, and student motivation. Preliminary analyses indicate that timing and format may combine to increase student achievement on the computer adaptive version. A controlled study was conducted in which two levels of time were crossed with two format presentations of reading text. The Nelson Denny Reading Comprehension Test was administered as an independent measure, and a questionnaire was also administered to assess student testing strategies relating to time and format.

Investigating Differences in Mean Score on Adaptive and Paper and Pencil Versions of the College Level Academic Skills Test

Abstract

This study examines the possible causes of a mean score increase for the computer adaptive form of a college level reading test over the traditional paper and pencil version. Issues addressed in the study include timing and format, measurement model procedures, content coverage and reading load, and student motivation. Preliminary analyses indicate that timing and format may combine to increase student achievement on the computer adaptive version. A controlled study was conducted in which two levels of time were crossed with two format presentations of reading text. The Nelson Denny Reading Comprehension Test was administered as an independent measure, and a questionnaire was also administered to assess student testing strategies relating to time and format.